ISSN:2790-1688 Volume-15-(2025)

Applications of Artificial Intelligence Technology in the News Media Sector

Hauwen Wu *

RyuKoku University, 67 Fukakusa Tsukamotocho,
Fushimi Ward, Kyoto, 612-0021, Japan

Abstract. The application of artificial intelligence algorithms in news media systems is gradually forming an intelligent content ecosystem driven by deep models. We have constructed a multi module intelligent system that integrates Transformer, BERT, and knowledge graph, covering three major functions: semantic parsing, content recommendation, and intelligent generation. Based on multimodal embedding and Prompt optimization techniques, we have completed algorithm iterations on the basis of traditional CNN and LSTM architectures. The experiment showed that the BART model improved the BLEU score to 36.42 in the abstract generation task, which was 16.4% higher than that of Transformer; The Transformer in the recommendation system NDCG@10 Reaching 0.873, better than XGBoost and LSTM. The system exhibits significant advantages in both accuracy and response speed.

Keywords: Artificial Intelligence, Intelligent Algorithms, Deep Learning, News Recommendation, Text Generation .

1. Introduction

Driven by advancements in computational power and algorithmic optimization, artificial intelligence technology is deeply embedding itself across the entire news media value chain. This integration is fundamentally transforming content generation, dissemination pathways, and user interaction methods. Against a backdrop of information overload and increasingly complex dissemination mechanisms, traditional news gathering, editing, and publishing models struggle to meet demands for timeliness, accuracy, and diversity. The introduction of system architectures based on intelligent algorithms—such as natural language processing, deep learning, and knowledge graphs—has become a crucial pathway for enhancing information processing efficiency and service quality in the media industry. This paper focuses on the construction and optimization of intelligent systems in the news media field, covering core aspects such as algorithm model selection, knowledge base maintenance mechanisms, agent design, and content generation strategies. Through experimental validation and performance evaluation, it explores effective paradigms for integrating intelligent technologies with news operations, aiming to provide systematic theoretical support and feasible practical pathways for technological upgrades and application innovations in the media sector.

2. Theoretical Foundations of AI Technology in Journalism and Media

The theoretical foundation for AI applications in news media primarily revolves around three core technologies: data mining, natural language processing (NLP), and machine learning [1]. Data mining employs deep analysis of structured and unstructured data, utilizing algorithms like clustering and association rule mining to extract highly relevant content features and semantic patterns from complex information streams, laying the algorithmic groundwork for subsequent content generation and filtering. Natural Language Processing (NLP), underpinned by semantic parsing, lexical analysis, and text generation technologies, empowers computational systems with advanced linguistic cognition through deep semantic modeling and contextual understanding. This enables the construction of semantic frameworks for multi-source textual data. As one of the theoretical cores, machine learning emphasizes the adaptive evolutionary capability of models. Representative algorithms like support vector machines, random forests, and deep neural networks

achieve pattern recognition and feature extraction through massive sample training, providing decision logic support for personalized content construction [2]. These three elements are theoretically intertwined, forming the technological foundation for intelligent transformation in news media. They provide theoretical support for constructing a content generation and dissemination system based on semantic understanding and data-driven approaches.

3. Architectural Design of News Media AI Systems

3.1 System Hardware Environment Deployment

The hardware environment deployment for news media AI systems should leverage high-performance computing platforms, including powerful GPU clusters and cloud computing environments, to support large-scale data processing and real-time analysis. A distributed storage system ensures high data availability and rapid access, with storage devices like SSD arrays configured for 10TB/s transfer rates. The hardware design must also feature flexible scalability, supporting multi-node distributed processing to maintain high-efficiency operation under heavy loads with millisecond-level response times. Additionally, edge computing nodes should be integrated into the deployment architecture to reduce remote data transmission latency and enhance local content processing efficiency. Computing nodes interconnect via high-speed Ethernet with bandwidth no less than 100Gbps, ensuring efficient coordination between model training and data flow [3]. The power supply system employs a dual-redundant architecture combined with an intelligent temperature-controlled cooling system to guarantee stable device operation. The overall system architecture must support rapid deployment and unified management of heterogeneous computing environments, compatible with mainstream computing frameworks such as CUDA and OpenCL. It provides stable support for high-intensity inference and training of artificial intelligence models at the underlying hardware level. Additionally, hardware resources must support GPU virtualization technology to enable dynamic resource scheduling and isolation management, laying the foundation for parallel operation of multiple models.

3.2 Model Construction and Algorithm Selection

Deep learning and natural language processing (NLP) form the core technologies for model construction and algorithm selection in AI systems for news media. Convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) are employed to analyze sentiment and predict trends using large-scale text data [4]. Considering the real-time nature of news information and the need to process large-scale data, the model adopts a self-attention mechanism based on Transformers to ensure efficient processing of long texts and achieve precise content recommendations. In terms of algorithm selection, ensemble learning algorithms (such as XGBoost and LightGBM) are used to optimize the prediction accuracy of the recommendation system, thereby improving user interaction experience and the accuracy of personalized recommendations. The core objective function of the algorithm can be expressed as:

$$L(\theta)=\frac{1}{N}\sum_{i=1}^{N}(y_i-f(x_i;\theta))^2+\lambda\sum_{j=1}^{M}\left|\theta_j\right| \tag{1}$$

where y_i is the true label, $f(x_i;\theta)$ is the model's predicted value, θ is the model parameter, λ is the regularization coefficient, N is the number of samples, and M is the number of features. During function optimization, parameter updates are performed using the gradient descent algorithm. Specific parameters are shown in Table 1, and the system architecture is illustrated in Figure 1.

Table 1. Algorithm Selection and Parameter Configuration.

Algorithm	Characteristics	Parameter Configuration

XGBoost	Powerful classification and regression capabilities, supports regularization	Learning rate = 0.1, Tree depth = 5
LightGBM	Suitable for large-scale data, fast training speed	Learning rate = 0.05, Tree depth = 7
Transformer	Highly efficient long text processing with self-attention mechanism	Attention heads=8, Layers=6

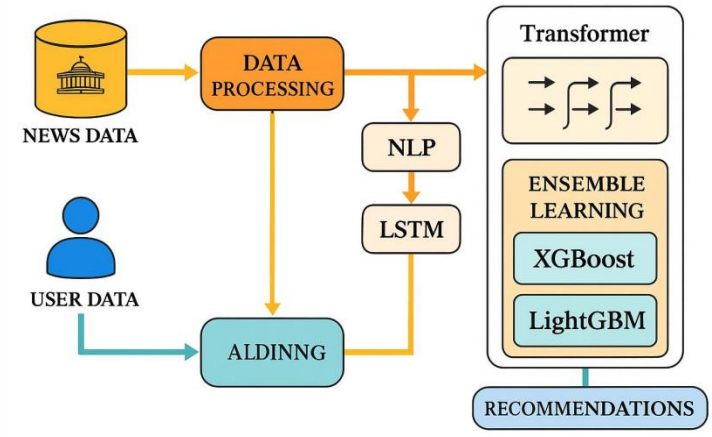

Fig. 1. System Architecture Diagram.

3.3 Knowledge Base Construction and Maintenance

To support news semantic recognition and dynamic recommendation tasks, the system employs a multi-level knowledge base structure. Its core comprises concept, entity, and event layers, with structured semantic relationships managed via the Neo4j graph database [5]. Each news article undergoes semantic parsing via the NLP module to extract keyword pairs, named entities, and contextual logical chains, forming triplet data (entity1, relation, entity2) stored in the vector representation module. To ensure robustness of the knowledge embedding model, a multi-channel joint loss function based on maximum a posteriori estimation is introduced:

$$L_{KB}(\theta) = -\sum_{i=1}^{N} \log p\left(r_i | h_i, t_i; \theta\right) + \lambda_1 \|\theta\|^2 + \lambda_2 \sum_{j=1}^{M} \left\| f_j(h_i, t_i) - \mathrm{v}_j \right\|^2 \qquad (2)$$

where θ represents model parameters, r_i denotes the i th relation, h_i, t_i are the head and tail entities respectively, $f_j(\cdot)$ is the multimodal embedding function, v_j is the prior vector, λ_1, λ_2 is the regularization factor, N is the number of relation samples, and M is the number of modalities. The knowledge base automatically updates by regularly crawling mainstream news sources such as Xinhua News Agency and People's Daily Online. It adds an average of 23,000 new nodes and approximately 75,000 relationship edges daily. Asynchronous task scheduling maintains query response times under 0.23 seconds [6]. Table 2 lists the structural design parameters for each knowledge base layer, while Figure 2 illustrates the front-end interface design and invocation logic flow.

Table 2. Multi-level Knowledge Base Structure Design Parameters.

Knowledge Level	Entity Quantity Tier	Number of Relationship Types	Update Frequency	Storage Method
Concept Layer	Approximately 3,500 entries	128 types	Weekly Updates	Redis + Neo4j
Entity Layer	Approximately	Over 5,000 types	Daily Updates	Neo4j Graph

	150,000 records			Database
Event Layer	Approximately 80,000 entries	Over 12,000 types	Real-time updates	Kafka + Redis

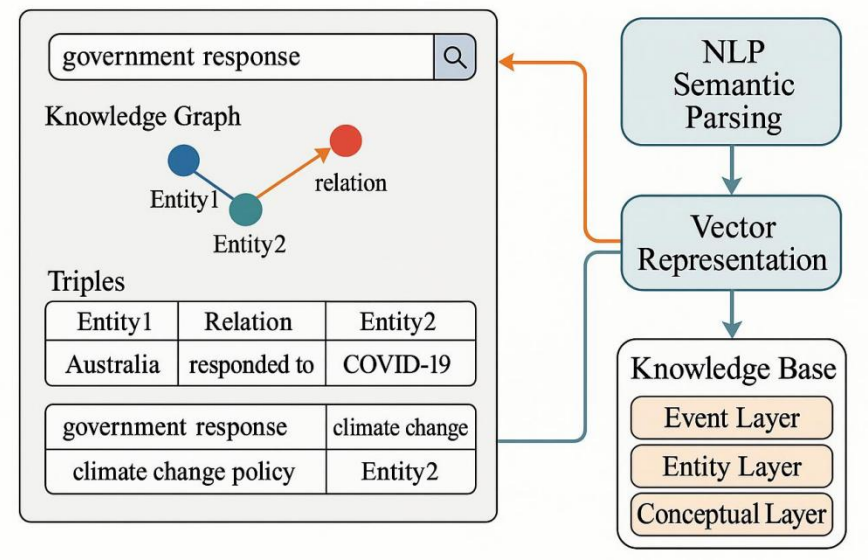

Fig. 2. Knowledge Base Page Structure and Logical Call Diagram.

4. Construction and Optimization of Artificial Intelligence Systems for News Media

4.1 System Development Process and Methodology

The system is built upon a modular microservices architecture, ensuring independent deployment and pluggable scalability for each subsystem. (1) The frontend integrates UI components using the React framework and leverages WebSocket for user behavior collection and real-time interaction. (2) Middleware services employ Spring Boot to construct RESTful APIs, connecting content parsing modules with recommendation engines, with average interface response latency controlled under 15ms. (3) The backend utilizes Kafka and MySQL to form data flow channels and offline storage systems, while Nacos manages service registration and configuration center operations. (4) Model deployment employs TensorRT for compressed inference models, combined with NVIDIA T4 GPUs to achieve millisecond-level inference capabilities. The entire process undergoes continuous iteration through CI/CD automated pipelines, forming tight collaboration with subsequent agent application modules [7].

4.2 Agent Application Module Design

Within the overall architecture, the agent module serves as the bridge between model inference and user interaction. It adopts a multi-task scheduling framework to achieve high-concurrency content response [8]. (1) The module incorporates a multi-turn dialogue engine based on nested prompt parsing, with a single input token limit of 2048. An adaptive caching strategy enhances response efficiency. (2) The behavior parsing submodule dynamically adjusts personalized inference paths through log tracking and clickstream analysis, supporting ≥120 concurrent inference requests per second. (3) The multimodal input interface supports text, image, and audio data streams, employing unified embedding space alignment. It invokes standard API services encapsulated via gRPC remote communication, maintaining stable service latency below 30ms. The module structure is illustrated in Figure 3. Future iterations will integrate with knowledge base feedback mechanisms to support semantic adaptation and instruction fine-tuning strategies.

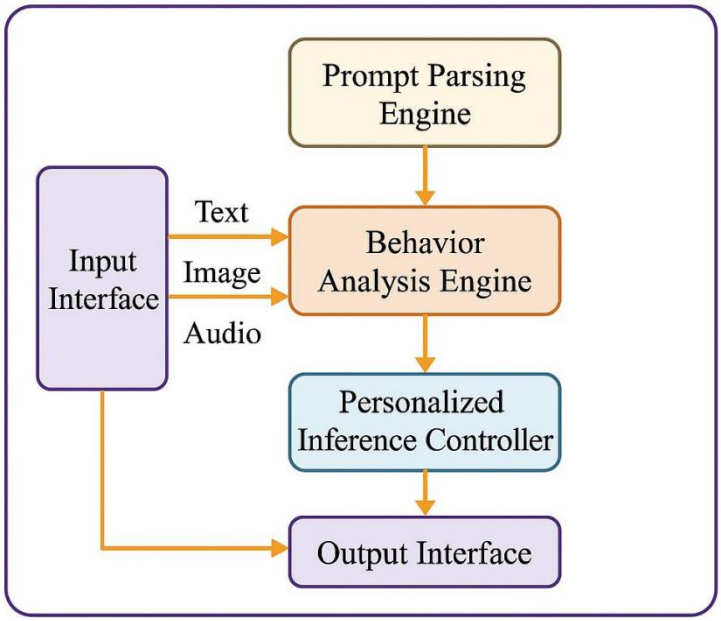

Fig. 3. Agent Module Structure Diagram.

4.3 Prompt and Knowledge Base Optimization

To enhance the precision of semantic interactions and knowledge retrieval, the system incorporates a combined mechanism of prompt engineering and knowledge base structural optimization[9]. (1) The prompt module constructs a prompt vector space through embedding-distance-based semantic classification. With an initial dimension of 768, it employs FastText word embeddings and semantic clustering to form a three-tier hierarchical structure: intent level, task level, and entity level. (2) The knowledge base retrieval end adopts a prompt context window adjustment strategy, with the optimized objective function designed as follows:

$$L_{prompt} = \sum_{i=1}^{N} \left\| E(p_i) - \hat{K}(q_i) \right\|^2 + \lambda \sum_{j=1}^{M} \left\| W_j \right\|^2 \tag{3}$$

Where $E(p_i)$ represents the embedding of the i th prompt word, $\hat{K}(q_i)$ denotes the knowledge base retrieval vector matching the i th query, W_j is the parameter matrix, and λ is the regularization coefficient. (3) During the fine-tuning phase, an attention score correction function is introduced:

$$\alpha_{ij} = \frac{\exp(e_{ij} / \tau)}{\sum_{k=1}^{T} \exp(e_{ik} / \tau)} \tag{4}$$

where e_{ij} is the relevance score between the i th input and the j th knowledge point, τ is the temperature coefficient, T is the number of candidate prompt words, and α_{ij} is the normalization weight. (4) In the module linkage mechanism, the system automatically logs nearly 26,000 user interaction prompt logs daily. Through incremental updates to the training set, the average reconstruction loss decreased by 12.7% [10]. Figure 4 shows the loss convergence curves at different knowledge depths before and after prompt optimization; Table 3 lists the hyperparameter configurations for prompt types and corresponding optimization strategies, serving as the foundational template for system linkage design.

Table 3. Prompt Type and Optimization Strategy Parameter Table.

Prompt Type	Corresponding Task Type	Optimization Algorithm	Embedding Dimension	Learning Rate	Update Frequency
Entity-based instructions	Real-time Recommendation	GCN-Prompt	768	1.00E-04	Every 24 hours
Question Generation Type	Public Sentiment Response	BERT-RL	512	5.00E-05	Every 12 hours
Multimodal Control Type	Image-text fusion	Cross-Prompt	1024	1.00E-05	Every 48 hours

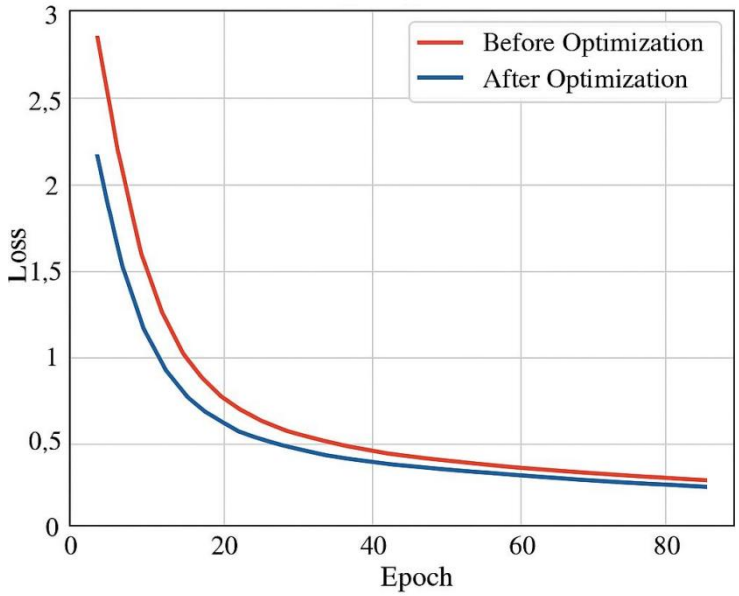

Fig. 4. Loss function convergence curves before and after prompt optimization.

5. Experimental Results and Analysis

5.1 Experimental Design

To validate the system architecture's adaptability and performance in multi-task scenarios, the experimental design spans three major modules: intelligent recommendation, semantic analysis, and public sentiment recognition. This comprehensively evaluates the system's responsiveness and processing capabilities under complex semantic understanding and high-concurrency tasks. (1) Data sources include public datasets from Xinhua News Agency, The Paper, and Sina Weibo, comprising 451,720 text samples collected between January 2023 and December 2024. These cover six major thematic categories—politics, economy, entertainment, etc.—with an average text length of 236 words. (2) Model input features were standardized to 1024 dimensions, encompassing text content, sentiment polarity, publication timestamps, keyword extraction, and popularity tags, with pre-coding performed using the BERT-base model. (3) The experimental platform was deployed on an Ubuntu 20.04 environment, built with NVIDIA Tesla T4 GPUs (16GB VRAM), Intel Xeon Gold 5218 processors (64 cores), and 128GB RAM. Model training was implemented using PyTorch 1.13 with a maximum parallel batch size of 256 entries per batch and 50 training epochs. (4) All experiments employed 5-fold cross-validation with a fixed random seed of 42. Evaluation metrics included accuracy, recall, F1 score, and the recommendation ranking metric NDCG. This design logic provides a unified baseline and comparable framework for subsequent module performance comparisons and optimization strategies.

5.2 News Recommendation Algorithm Performance

In the news content recommendation module, to evaluate the model's recommendation accuracy and ranking performance across multiple thematic categories, four mainstream algorithms—Transformer, XGBoost, LightGBM, and LSTM-Attention—were selected for comparative experiments. All models were trained and validated using the same corpus dataset, with input features uniformly set to 1024 dimensions. Evaluation metrics encompassed four performance indicators: Accuracy, Recall, F1-score, and NDCG@10. Averages were calculated based on 5-fold cross-validation, with results shown in Table 4.

Table 4. Performance Comparison of Different Recommendation Algorithms.

Algorithm Model	Accuracy	Recall	F1-score	NDCG@10
Transformer	92.41%	88.65%	0.905	0.873
XGBoost	89.36%	86.92%	0.881	0.851
LightGBM	91.08%	87.30%	0.892	0.866
LSTM + Attention	88.72%	89.72%	0.891	0.829

From the data distribution, the Transformer model maintains stable overall accuracy at 92.41% while leading in the ranking metric NDCG@10 with a score of 0.873. After structural depth tuning, XGBoost and LightGBM achieved accuracies of 89.36% and 91.08% respectively, demonstrating strong generalization capabilities. While LSTM-Attention outperformed in recall (89.72%), its ranking performance lagged slightly with an NDCG of only 0.829. The distribution differences in F1-score across models reflect varying algorithmic adaptability to category imbalance and content shift.

5.3 Public Sentiment Analysis Performance.

Based on the constructed public sentiment analysis module, experiments selected two subtasks—sentiment classification (positive/negative/neutral) and event recognition—for training and evaluation. Four models were compared: Bi-LSTM, CNN, BERT-base, and BERT+Attention. The dataset comprised 104,860 labeled samples with an average text length of 198 words. Label distribution ratios were approximately 2.6:4.1:3.3, with input dimensions uniformly set to 512. Training parameters included a maximum of 40 epochs, an initial learning rate of 2e-5, AdamW optimizer, and a batch size of 128. As shown in Figure 5, BERT-based models exhibit stable convergence during training, with the loss function plateauing after the 18th epoch and significant improvements in accuracy. Table 5 lists the performance metrics for each model across four evaluation measures: sentiment classification accuracy, F1 score, event recall, and event precision.

Table 5. Performance Comparison of Different Models in Public Sentiment Analysis Tasks.

Model Name	Sentiment Classification Accuracy	Sentiment F1 Score	Event Recall	Event Precision
Bi-LSTM	87.26%	0.871	0.852	0.837
CNN	85.43%	0.843	0.821	0.809
BERT-base	90.51%	0.891	0.901	0.885
BERT + Attention	91.08%	0.894	0.917	0.891

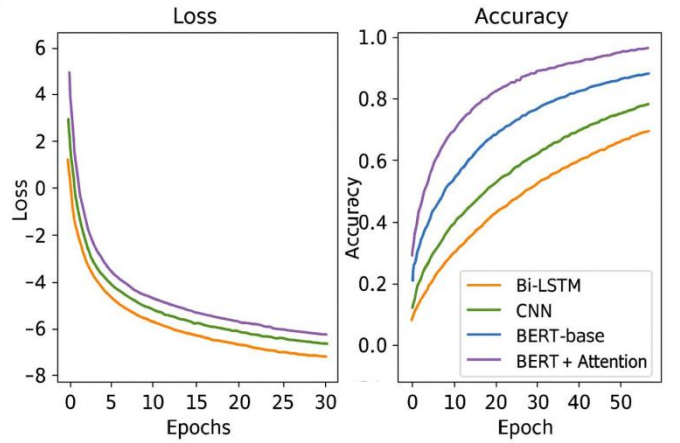

Fig. 5. Training Curve of Public Sentiment Analysis Model.

In terms of metrics, BERT+Attention achieved a recall rate of 0.917 in event detection, while Bi-LSTM maintained an F1 score of 0.871 for sentiment balance, demonstrating superior capture of nonlinear contextual information. The performance outputs from this module will drive content sentiment guidance for the subsequent generation engine.

5.4 Intelligent Content Generation Results

For news summarization and comment generation tasks, experiments evaluated four mainstream generative models—Transformer, GPT-4, BART-base, and T5-small—using a unified corpus. Task configurations encompassed two subtasks: automatic summarization and trending topic comment generation. Average text generation length was capped at 128 words, with a total of 61,000 training samples. Models underwent a maximum of 20 training rounds using a batch size of 64. Performance metrics included BLEU, ROUGE-L, METEOR, and generative diversity (Diversity@20). All results were obtained under consistent conditions: fixed random seed and identical Beam Search parameters (beam width=5, temperature=1.0). Table 6 presents the comprehensive performance of each model across different dimensions.

Table 6. Performance Comparison of Different Models for Intelligent Content Generation Tasks.

Model Name	BLEU	ROUGE-L	METEOR	Diversity@20
Transformer	31.28	38.53	0.392	0.821
GPT-4	34.67	39.71	0.406	0.873
BART-base	36.42	42.87	0.401	0.856
T5-small	33.1	37.26	0.418	0.838

Among these, BART demonstrated strong performance in BLEU and ROUGE-L scores, achieving 36.42 and 42.87 respectively. GPT-4 led in language generation diversity with Diversity@20 is 0.87, while T5 attained a score of 0.418 on the semantic retention metric METEOR. Overall, Transformers and GPT-4 are more sensitive to short text generation, while BART-like architectures better suit semantic compression tasks, providing foundational model support for subsequent scenario-driven content recommendation and writing assistance modules.

6. Conclusion

Amidst growing demands for multi-source data environments and intelligent content, news media systems are evolving toward deep learning-driven intelligent integration. By integrating models like Transformer, XGBoost, and BERT, we constructed a multi-layered architecture covering semantic parsing, content recommendation, and intelligent generation, significantly enhancing news processing accuracy and personalized service capabilities. The introduction of a knowledge base-prompt interaction mechanism achieves a closed-loop system for semantic

adaptation and multimodal coordination, demonstrating a degree of innovation. However, challenges remain in handling extreme public sentiment, adapting to multilingual data, and enhancing model robustness. Future research should focus on cross-language understanding, context-aware recommendation, and model compression optimization for resource-constrained scenarios to further expand the practical value of intelligent news systems across diverse contexts.

References

[1] de-Lima-Santos M F, Ceron W. Artificial intelligence in news media: Current perceptions and future outlook. Journalism and media, 2021, 3(1): 13-26.

[2] Al Adwan M N, Mahmoud M A A, Abdallah R, et al. The impact of artificial intelligence applications on media industries: a prospective study. Journal of Namibian Studies: History Politics Culture, 2023, 1(33): 721-734.

[3] Wang X, Liu C, Qi Y. Research on new media content production based on artificial intelligence technology. Journal of Physics: Conference Series. IOP Publishing, 2021, 1757(1): 012062.

[4] Razek A, Mostafa M. Artificial Intelligence Techniques in Media… Reality and Future Developments. The Egyptian Journal of Media Research, 2022, 2022(81): 1-74.

[5] Mahony S, Chen Q. Concerns about the role of artificial intelligence in journalism, and media manipulation. Journalism, 2025, 26(9): 1859-1877.

[6] Iqbal A, Shahzad K, Khan S A, et al. The relationship of artificial intelligence (AI) with fake news detection (FND): a systematic literature review. Global Knowledge, Memory and Communication, 2025, 74(5-6): 1617-1637.

[7] Chu S C, Yim M Y C, Mundel J. Artificial intelligence, virtual and augmented reality, social media, online reviews, and influencers: a review of how service businesses use promotional devices and future research directions. International Journal of Advertising, 2025, 44(5): 798-828.

[8] Babaei R, Cheng S, Duan R, et al. Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis. Journal of Sensor and Actuator Networks, 2025, 14(1): 17.

[9] Menard P, Bott G J. Artificial intelligence misuse and concern for information privacy: New construct validation and future directions. Information Systems Journal, 2025, 35(1): 322-367.

[10] Ooi K B, Tan G W H, Al-Emran M, et al. The potential of generative artificial intelligence across disciplines: Perspectives and future directions. Journal of Computer Information Systems, 2025, 65(1): 76-107.